ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

SIB Swiss Institute of Bioinformatics

Biophysical modeling of transcription factor binding sites using large SELEX libraries and computational simulations

Workshop on Bioinformatics of Gene Regulation

on the occasion of

30 Years TRANSFAC

Göttingen 7.-9. March 2018

Philipp Bucher

Transcription Factor

Original Definition:

Factors (proteins ?) necessary for transcription that are not part of (do not co-purify with) RNA polymerase

Today:

Gene regulatory proteins (in a broad sense) that interact with DNA or chromatin.

Two classes:

- Sequence specific DNA binding, *e.g.* CTCF, AP-1
- Others: EP300, Suz12 (not relevant for this talk)

More about transcription factor binding sites (TFBS)

Properties:

- High degeneracy: many related sequences bind same TF, *e*.g. TATAAA, TTTAAA, TATAAG, TTTAAG, etc.
- Short length: 6-20 bp
- Low specificity: 1 site per 250 to 25000 bp

Binding mode:

- Many factors bind as obligatory dimers or multimers
- Quantitative recognition mechanism: affinity of different binding sequences varies (affinity = DNA-protein binding equilibrium constant K_b, unit: Mol^{-1}, low values mean high affinity).
- Regulatory function often depends on cooperative interactions with neighboring TFs/sites (combinatorial gene regulatory code).

Formal Tools to Describe TF binding Motifs: Consensus Sequences and Position Weight Matrices

Consensus sequences:

- example: **TATAWA** (for eukaryotic TATA-box)
- a limited number of mismatches may be allowed
- may contain IUPAC codes for ambiguous positions, *e.g.* W = A or T.

Position Weight Matrices (PWM):

- a table with numbers for each residue at each position of the motif

Pos.	1	2	3	4	5	6	7	8	9
A:	6	10	1	0	21	92	15	2	6
C:	78	5	0	1	8	0	1	51	9
G:	12	0	1	4	66	2	1	44	6
T:	4	85	98	95	5	6	83	3	79

Many synonyms in use: Position-Specific Scoring Matrix (PSSM), Position Frequency Matrix (PFM), Base Probability Matrix (BPM), etc.

Two Major PWM Types: Frequency and Scoring Matrices

Frequency matrices directly reflect the relative frequencies of the four bases at consecutive motif positions

Scoring matrices contain numbers that are used to score DNA *k*-mers (sequences of same length as motif).

Position frequency matrix (horizontal)

6	10	1	0	21	92	15	2	6
78	5	0	1	8	0	1	51	9
12	0	1	4	66	2	1	44	6
4	85	98	95	5	6	83	3	79

Integer scoring-matrix (horizontal)

-6	-4	-11	-14	-1	6	-2	-9	-6
5	-6	-14	-11	-5	-14	-11	3	-4
-3	-14	-11	-7	4	-9	-11	2	-6
-7	5	6	6	-6	-6	5	-8	5

Base probability matrix (vertical)

0.06	0.78	0.12	0.04
0.10	0.05	0.00	0.85
0.01	0.00	0.01	0.98
0.00	0.01	0.04	0.95
0.21	0.08	0.66	0.05
0.92	0.00	0.02	0.06
0.15	0.01	0.01	0.83
0.02	0.51	0.44	0.03
0.06	0.09	0.06	0.79

A base probability matrix defines a motif as a:

Probability distribution over *k*-mers

A scoring matrix together with a cut-off value defines a motif as a:

Subset of all *k*-mers

Inference of PWM models

Source data:

- Sets of putative binding sequences defined/obtained by
  - *in vivo*: footprints, ChIP(-seq)
  - *in vitro*: bandshifts (EMSA), SELEX
- Quantitative affinity measurements of selected oligonucleotides
  - EMSA competition assays
  - Protein-binding microarrays (PBMs)

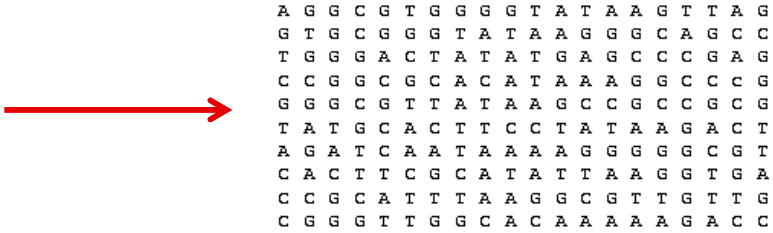
Computational motif inference:

- Motif discovery algorithms (for sequence sets)
- Specialized parameter fitting algorithms for quantitative data

Important: Model quality depends on data quality and computational inference procedure (the latter may be more critical)

Motif Discovery Overview

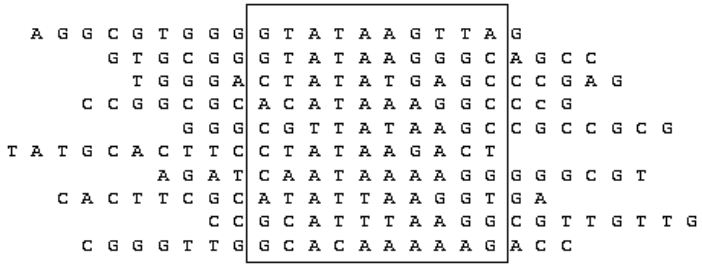
Input sequences longer than motif, motif positions unknown.



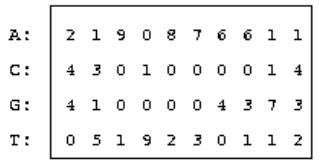
Motif positions inferred (guessed) by some kind of algorithm:

- Word search algorithms
- Iterative alignment, EM

Re-alignment of sequences

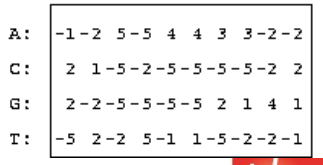


Position frequency matrix



(converted into)

Log-odds (weight) matrix



About SELEX

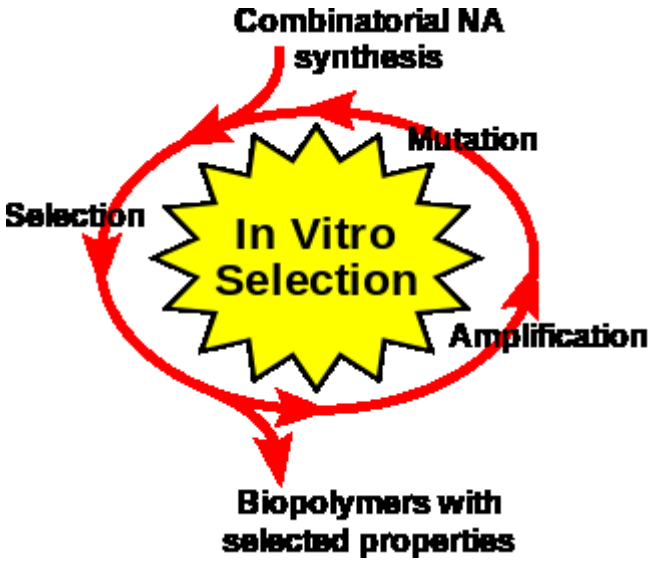
Systematic **E**volution of **L**igands by **EX**ponential Enrichment

Purpose:

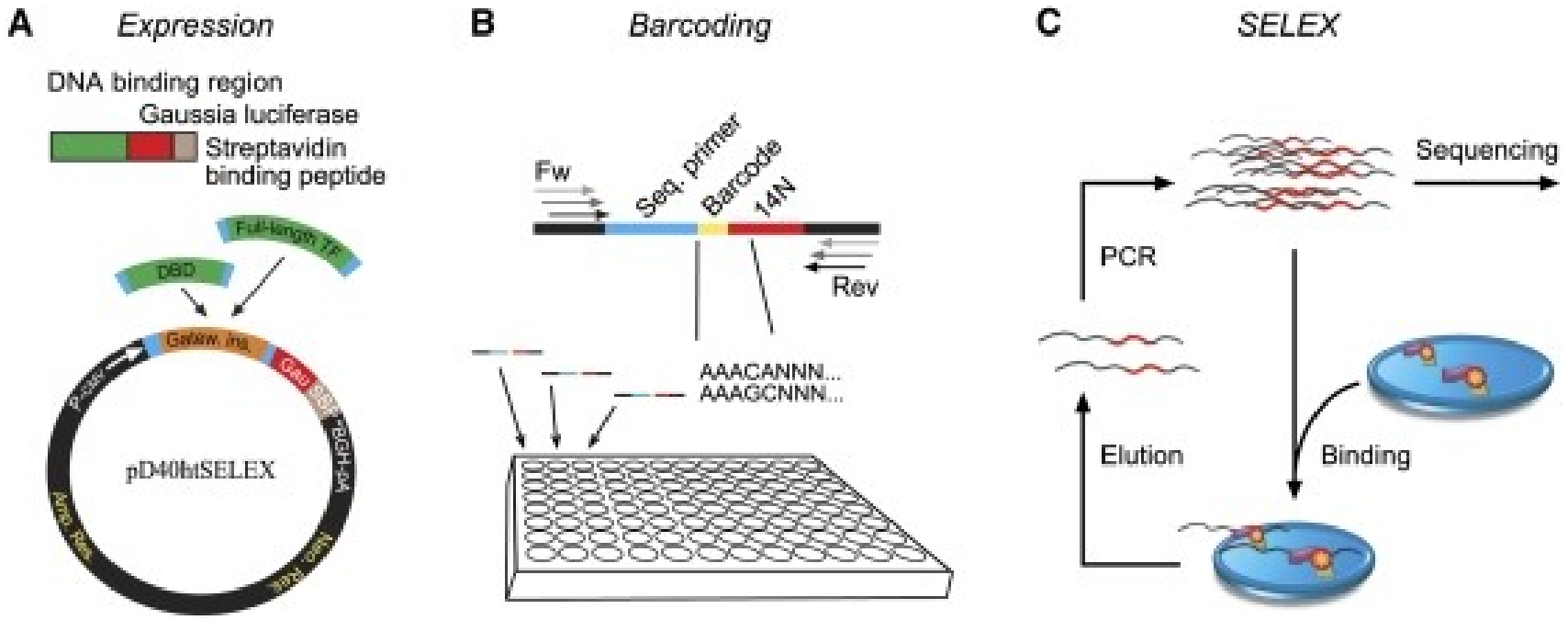
- To generate high-affinity nucleic acid ligands to be used as drugs or reagents (e.g. aptamers)
- Comprehensive characterization of the binding specificity of DNA or RNA binding proteins

Selection technique for TF ligands:

- Affinity chromatography
- Gel shifts (Roulet et al. Nature Biotechnol 2002)
- Immobilized proteins on 96 well plates (Jolma et al. Genome Res 2010)
- Microfluidic devices SMilE-seq (Isakova et al. Nat Methods 2017)



Example of a high-throughput SELEX protocol



Yield: up to 500'000 sequences per library

Jolma et al. 2010. Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res. 20:861.

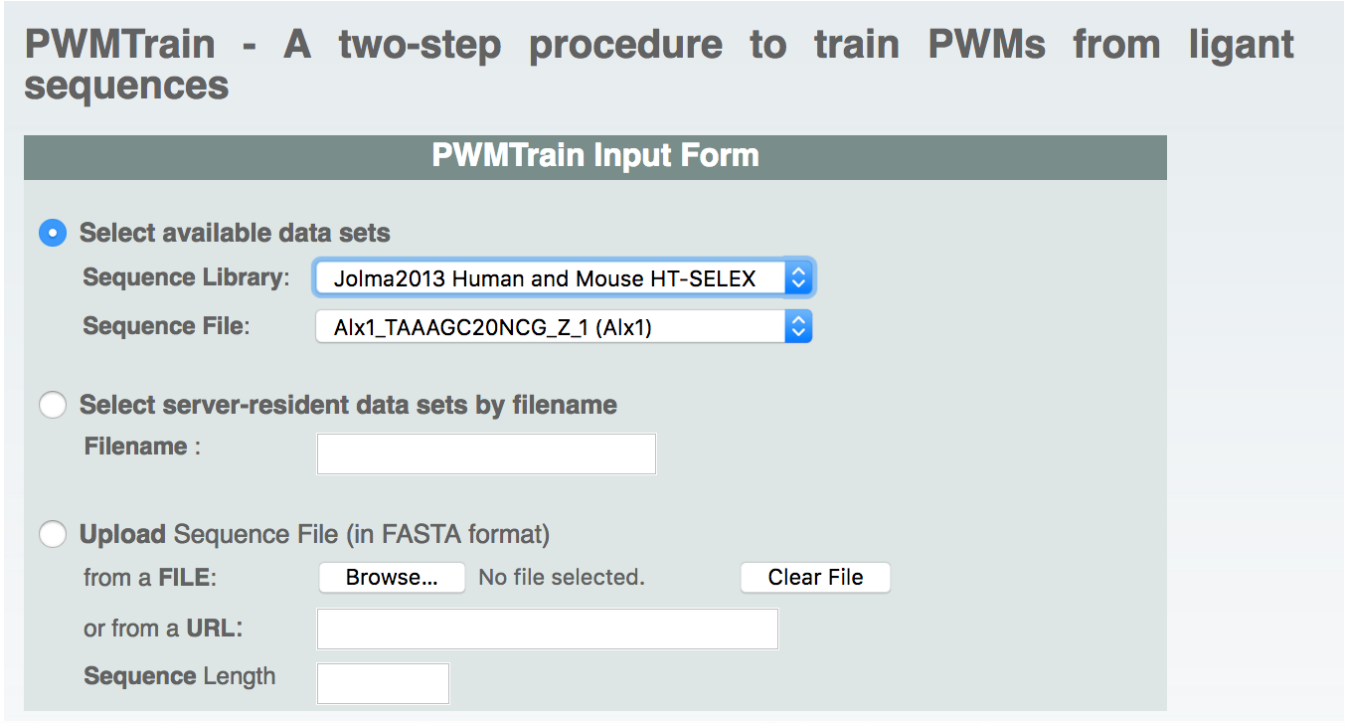
Our PWM inference method for SELEX data

Find suitable over-represented *k*-mer with word search algorithm

Optional: extend *k*-mer consensus sequence by few insignificant positions (Ns)

Optimize consensus-derived PWM using EM via a hidden Mark model

Reference: Isakova *et al.* 2017, Nat Methods. 14(3):316-322.
Web server: http://ccg.vital-it.ch/pwmtools/pwmtrain.php



Word Search Algorithm Example: Herpes simplex Virus Promoters

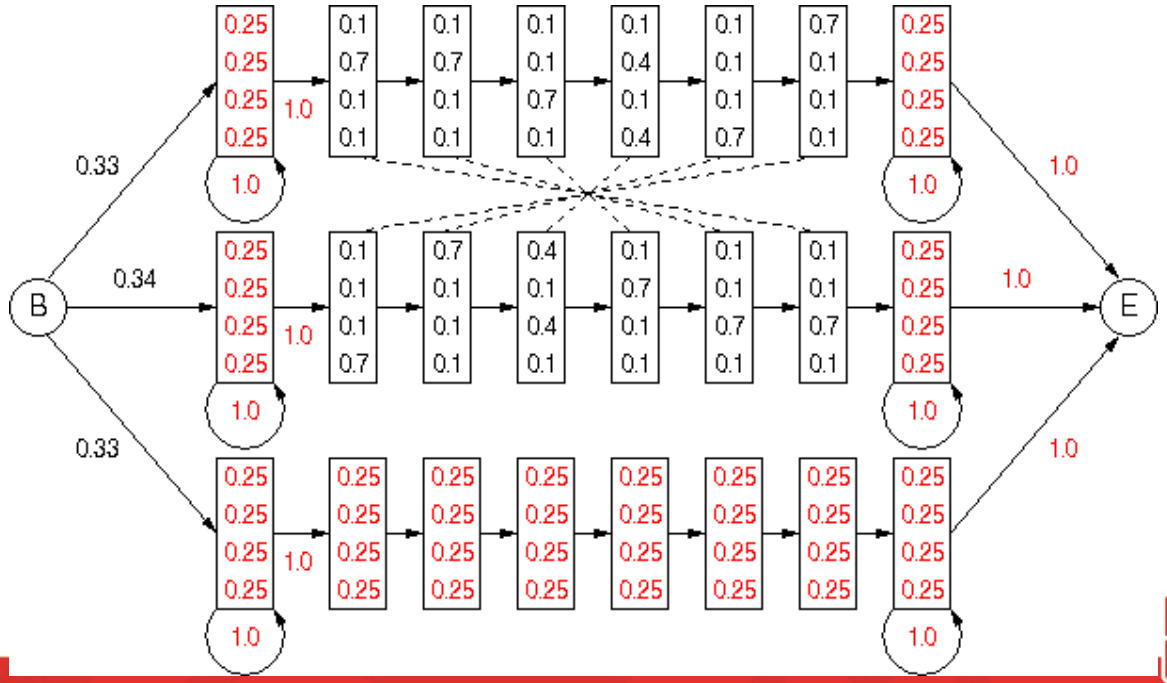
Pos. relative	-30 '	-20 '
HSV-1 IE-I	AGGCGTGGGGTATAAG	
HSV-1 IE-II	CCACGGGTATAAGGAC	
HSV-1 IE-III	TGGGACTATATGAGCC	
HSV-1 IE-IV/V	CCGGCGCACATAAAGG	
HSV-1 b' 82K AlkExo	GCTTAAGCTCGGGAGG	
HSV-1 b' 42K	TATGCACTTCCTATAA	
HSV-1 b' 39K dUTPase	CACACGCCCATCGAGG	
HSV-1 b' 33K	GATGTTTACTTAAAAG	
HSV-1 b' 21K	AGATCAATAAAAGGGG	
HSV-1 b' 5 kb	GATGTGGATAAAAAGC	
HSV-1 b' RNR2	TCCACGCATATAAGCG	
HSV-1 b' tk	CACTTCGCATATTAAG	
HSV-1 b' dbp	GTAAAGTGTACATATA	
HSV-1 b' gB 3.3 kb	GCCTGGCGATATATTC	
HSV-1 b' gD	GTCTGTCTTTAAAAAG	
HSV-1 b' gE	GCGCATTTAAGGCGTT	
HSV-1 b' ICP 18.5	CATCCGTGCTTGTTTG	
HSV-1[U-S] b' tr-4	CGGGTTGGCACAAAAA	
HSV-1[U-S] b' tr-9	CCGAGGCGCATAAAGG	
HSV-1 b'g' VP5	GGGGGGGTATATAAGG	
HSV-1 b'g' 2.1 kb	ACGTGATCAGCACGCC	
HSV-1 b'g' a'TIF/VSP	GGGTTGCTTAAATGCG	
HSV-1 b'g' 2.7 kb	CTCCTCCCGATAAAAA	
HSV-1 g' 5 kb	GGCCCGCGTATAAAGG	
HSV-1 g' gC	CCCGGGTATAAATTCC	
HSV-1 g' gH	CAGAATAAAACGCACG	
HSV-1 g' 42K	AACCTTCGGCATAAAA	
HSV-1 Ori_s ORF	GTGCGTCCCCTGTGTT	
HSV-1 18K	GGCGCTATAAAGCCGC	

Word	Frequency	Enrichment	Log(P-val)
ATAAA	10	4.4894288	-10.718707
TATAA	8	4.5869487	-9.351983
GATCA	2	19.5289309	-8.704620
CGCAT	4	7.7738351	-8.535924
ACTTC	2	14.2323077	-7.783882
GTATA	5	5.2344852	-7.665632
GCACA	2	13.4990797	-7.630886
CACTT	2	12.3479312	-7.373765
CGAGG	2	12.2250286	-7.344967
CTTCG	2	12.2121607	-7.341936
CACGC	3	7.4119396	-7.117540
GCATA	4	5.5593843	-7.027697
CCACG	2	10.9045829	-7.016787
GATGT	2	10.8879457	-7.012415
AAAGG	4	5.4604691	-6.948596
TAAAG	5	4.4597446	-6.843935
TGTTT	2	9.6585434	-6.670336
TAAAA	7	3.3983314	-6.631400
AGGCG	3	6.4831545	-6.627692
CTTAA	3	6.1010158	-6.407487
GGTAT	4	4.5535294	-6.159526
TTAAG	3	5.5955386	-6.096445
CGCAC	2	7.8370370	-6.079061
GGGTA	4	4.3205355	-5.935471
AGGAC	1	13.2320762	-5.908658

HMM-based method for PWM construction

Principle:

- Model SELEX sequences (binding sites plus flanks or background) with a hidden Markov model (HMM)
- Define an initial model with consensus sequence like binding site
- Train with EM, extract binding site model from EM.



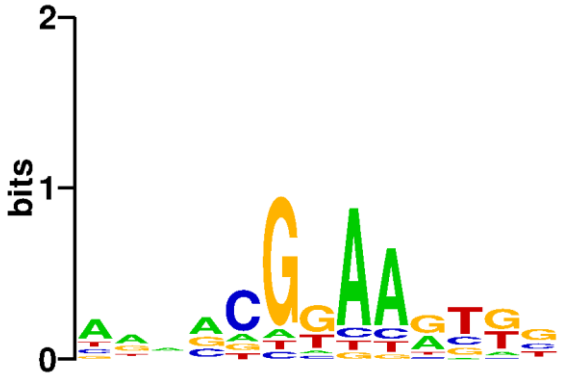
Models from later SELEX cycles get more skewed.

Example: ELF3_TCCGTG20NTGC_Y (seed NNNCCGGAAGNNN)

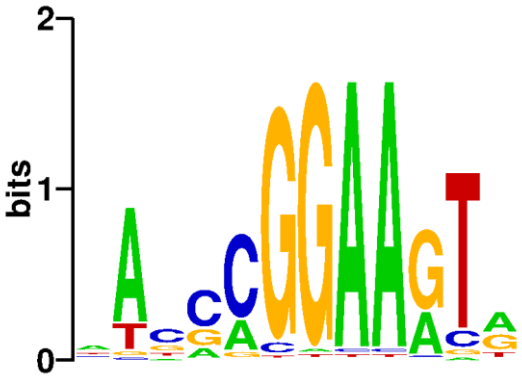
Which one is the correct model?

Are the differences relevant?

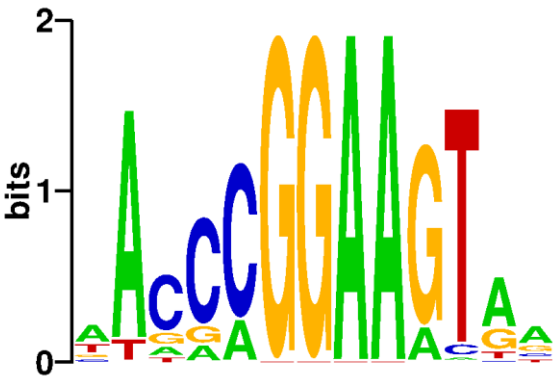
Cycle 1



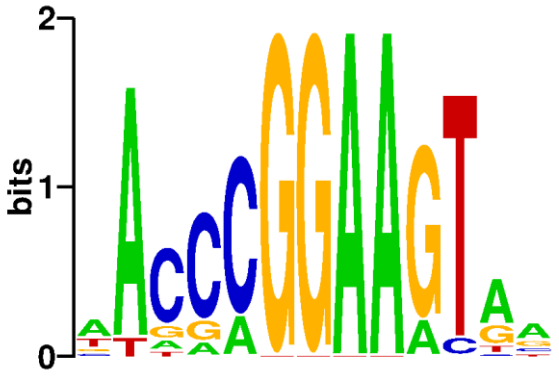
Cycle 2



Cycle 3



Cycle 4



Models from later SELEX get more skewed

ELF3_TCCGTG20NTGC_Y cycle 2

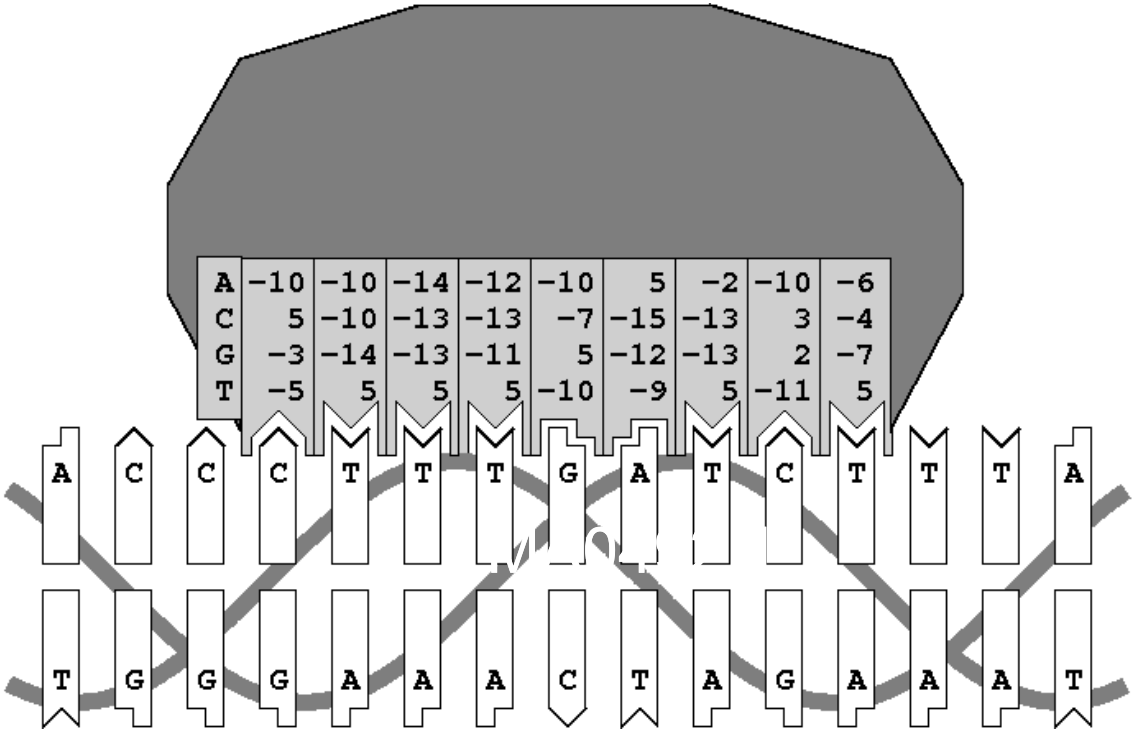
0.417	0.188	0.162	0.233
0.759	0.023	0.041	0.177
0.137	0.483	0.229	0.151
0.177	0.537	0.266	0.020
0.254	0.669	0.061	0.015
0.013	0.048	0.924	0.015
0.021	0.012	0.954	0.013
0.959	0.014	0.008	0.019
0.951	0.022	0.008	0.019
0.333	0.030	0.628	0.010
0.035	0.099	0.045	0.820
0.441	0.068	0.347	0.144
0.334	0.202	0.262	0.202

ELF3_TCCGTG20NTGC_Y cycle 3

0.488	0.119	0.133	0.260
0.906	0.002	0.004	0.088
0.124	0.655	0.142	0.078
0.122	0.741	0.137	0.001
0.212	0.784	0.002	0.001
0.002	0.001	0.996	0.001
0.002	0.001	0.996	0.001
0.997	0.001	0.001	0.001
0.992	0.002	0.001	0.004
0.163	0.004	0.832	0.001
0.024	0.055	0.004	0.916
0.593	0.040	0.258	0.109
0.502	0.143	0.216	0.139

Red: preferred base, blue: least preferred base

Physical Interpretation of Transcription Factor PWM



Weight matrix elements represent relative binding energies between DNA base-pairs and protein surface areas (base-pair acceptor sites).

A weight matrix column describes the base preferences of a base-pair acceptor site.

Berg-von Hippel model of protein-DNA interactions

The weight matrix score expresses the binding free energy of a protein-DNA complex in arbitrary units:

$$-\Delta G(x) = S(\mathbf{x}) + \text{const.}$$

$$S(\mathbf{x}) = \sum_{i=1}^{N} w_i(x_i)$$

It is convenient to express the binding free energy in dimension-less *RT* units:

$$E(\mathbf{x}) = \sum_{i=1}^{N} \varepsilon_i(x_i)$$

$$\varepsilon_i(b) = -w_i k(b)$$

On a relative scale, the binding constant for sequence **x** is then given by:

$$K_{\text{rel}}(\mathbf{x}) = e^{E(\mathbf{x})}$$

For sequences longer than the weight matrix, we may use:

$$K_{\text{rel}}(\mathbf{x}) = \frac{1}{\sum_i e^{-E(x_i \ldots x_{i+N-1})}} \quad \text{or} \quad K_{\text{rel}}(\mathbf{x}) = \frac{1}{\max_i e^{-E(x_i \ldots x_{i+N-1})}}$$

(index i runs over all subsequence starting positions on both strands)

Berg-von Hippel Theory – the λ parameter

The energy terms of a weight matrix can be computed from the base frequencies $p_i(b)$ estimated from *in vitro* or *in vivo* selected binding sites:

$$\varepsilon_i(b) = -\frac{1}{\lambda} \ln \frac{p_i(b)}{q(b)}$$

$q(b)$ is the background frequency of base b.

λ is an unknown parameters related to the stringency of the binding conditions.

The probability that a specific DNA sequence is bound depends on the "chemical potential" μ which is a function of the relative protein and DNA concentrations.

$$P(bound \mid S_j) = \frac{e^{-E_j}}{e^{-\mu} + e^{-E_j}} = \frac{1}{1 + e^{E_j - \mu}}$$

Note: μ is the energy of a DNA sequences which has binding probability 0.5.

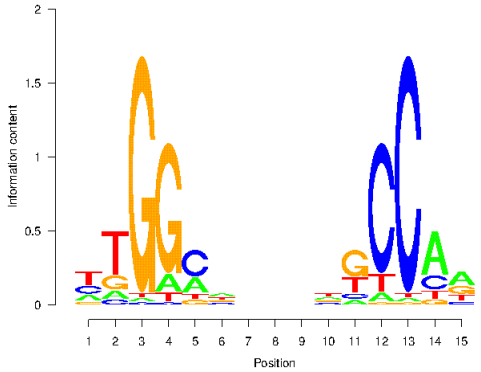
Estimating the λ parameter from absolute binding constants
Example CTF/NF1

Experimental data from Meisterernst et al. (1988). *Nucleic Acids Res.* 16, 4419-4435.

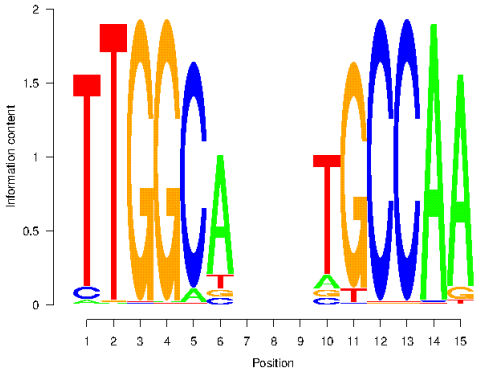
PWM from Roulet et al. 2002, *Nat. Biotechnol.* 20, 831-835

4	-2	-20	-4	6	0	0	0	0	-3	-2	-9	-20	10	10
6	-5	-23	-19	10	-4	0	0	0	-4	-2	10	10	2	-1
-1	2	10	10	-2	-4	0	0	0	-4	-10	-19	-23	-5	6
10	10	-20	-9	-2	-3	0	0	0	0	6	-4	-20	-2	4

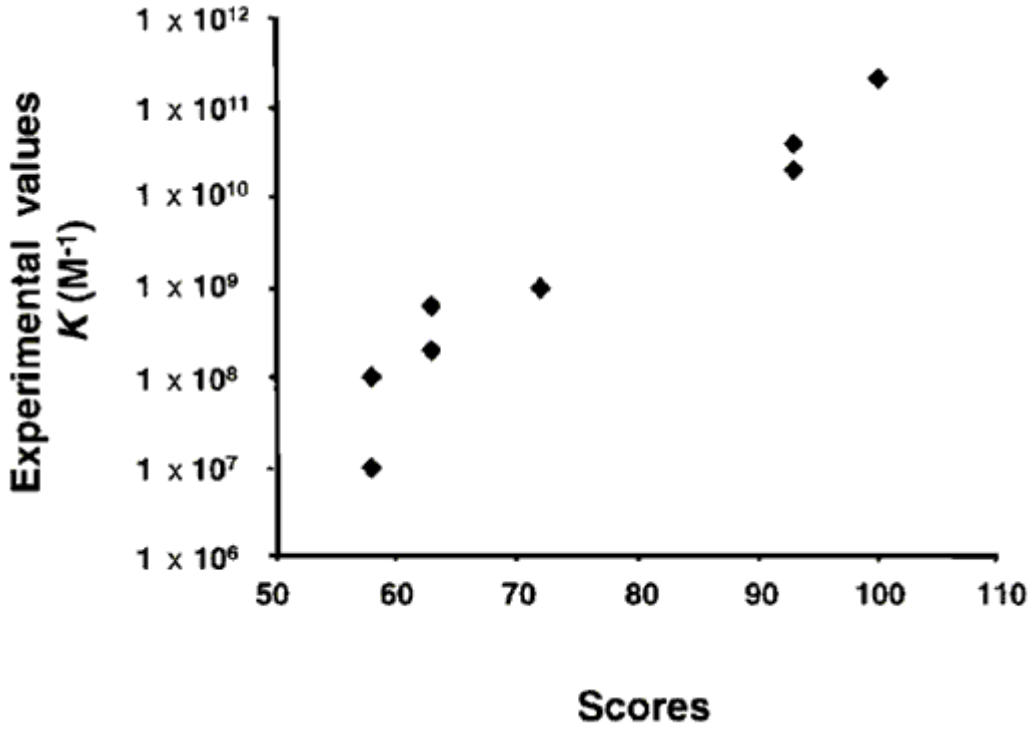
λ =0.100



λ =0.500



Example: Fitting weight matrix scores of CTF/FN1 binding site to absolute binding constants



Experimental data from Meisterernst et al. (1988). *Nucleic Acids Res.* 16, 4419-4435.

Predicted binding scores from Roulet et al. 2002, *Nat. Biotechnol.* 20, 831-835

A factor of 10 in K values corresponds to a difference of ~12 in weight matrix score values → $\lambda \approx \ln(10)/12 = 0.192$.
(relative binding energies estimated by best susbsequence score)

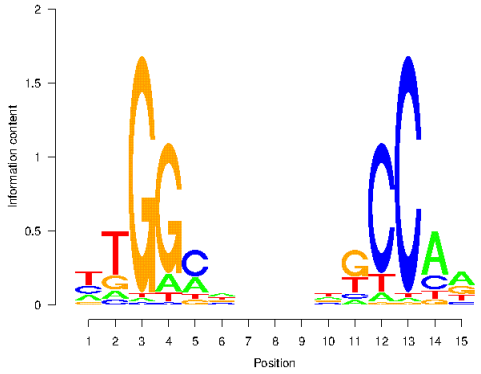
Estimating the λ parameter from absolute binding constants
Example CTF/NF1

Experimental data from Meisterernst et al. (1988). *Nucleic Acids Res.* 16, 4419-4435.

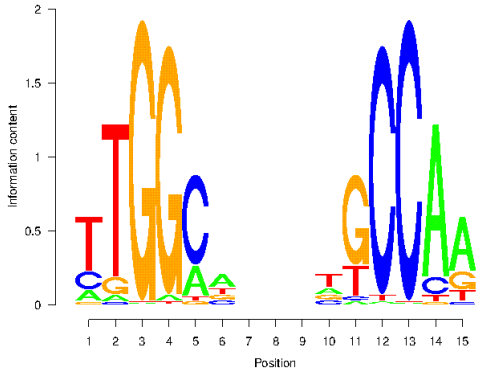
PWM from Roulet et al. 2002, *Nat. Biotechnol.* 20, 831-835

4	-2	-20	-4	6	0	0	0	0	-3	-2	-9	-20	10	10
6	-5	-23	-19	10	-4	0	0	0	-4	-2	10	10	2	-1
-1	2	10	10	-2	-4	0	0	0	-4	-10	-19	-23	-5	6
10	10	-20	-9	-2	-3	0	0	0	0	6	-4	-20	-2	4

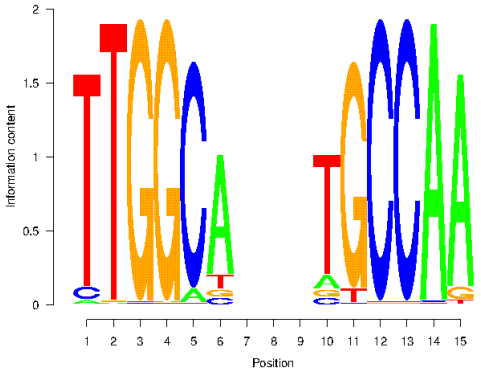
λ =0.100



λ =0.192



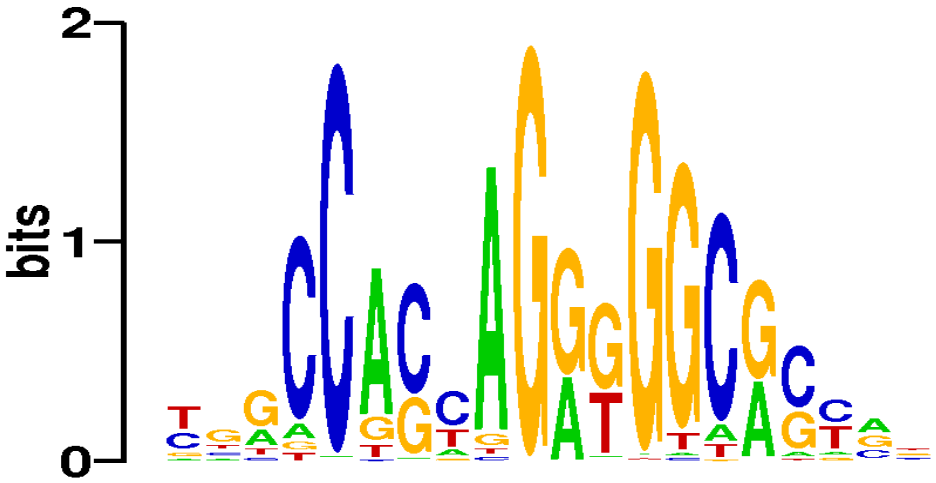
λ =0.500



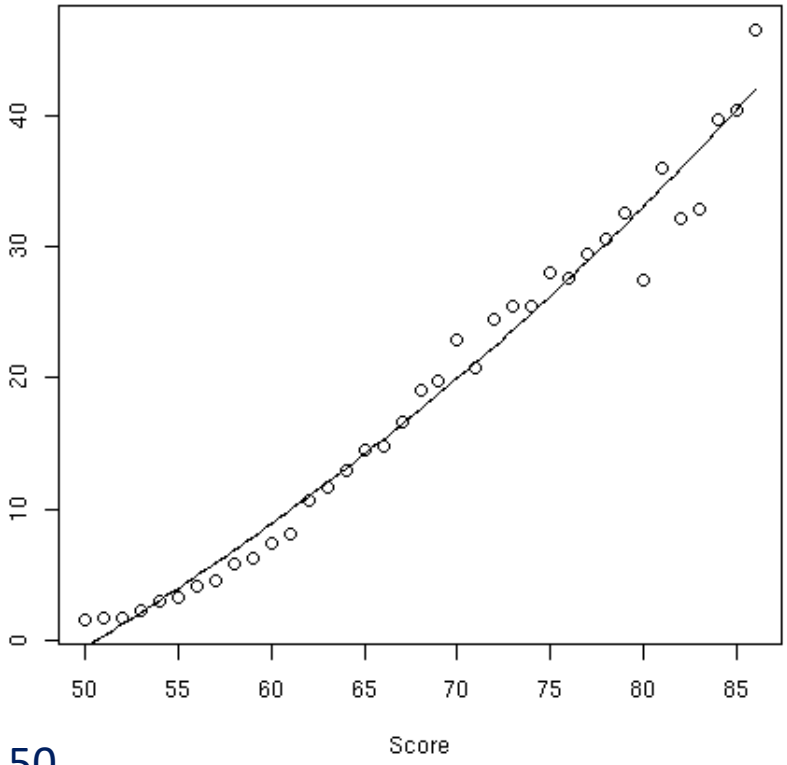
The "*in vivo*" λ could be estimated from ChIP-seq tag coverage

Example CTCF sites (Chip-seq data from Barski *et al.* 2007)

CTCF weight matrix (Sequence logo):



Average # of counts per predicted site



Matrix derived from 3888 highly enriched peaks
Maximal score of weigh matrix: 90
Threshold score: 50
Human genome contains 77924 sites with score ≥ 50
Tags counted in window of 100 bp

Next Step: modeling the entire experiment rather than just the binding sites

Motivation:

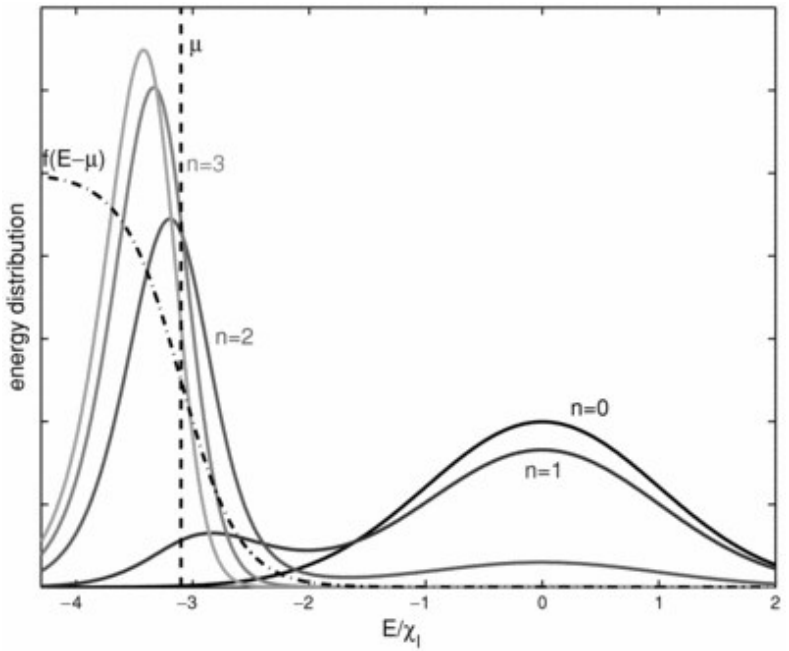
- Current motif discovery algorithm optimize a sequence-intrinsic quality measure such as information content or Maximum-Likelihood
- These measures are not justified by a biophysical model of the SELEX process
- Modeling additional parameters of the experiment such as the chemical potential and non-specific binding is necessary to obtain a biophysical model that completely explains the data
- An accurate model of a SELEX experiment should reproduce all statistical properties of the SELEX library (*e.g. k*-mer counts) in a computational simulation

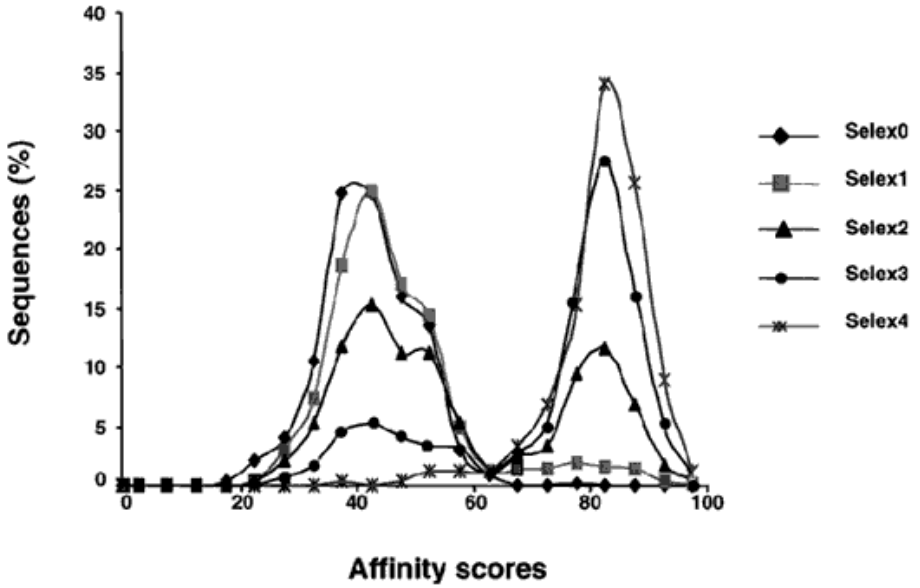
The PWM score distributions of SELEX libraries selected under constant chemical potential are predicted by a biophysical model

Theoretically predicted affinity profiles of successive SELEX cycles (Djordjevic & Sengupta 2006)

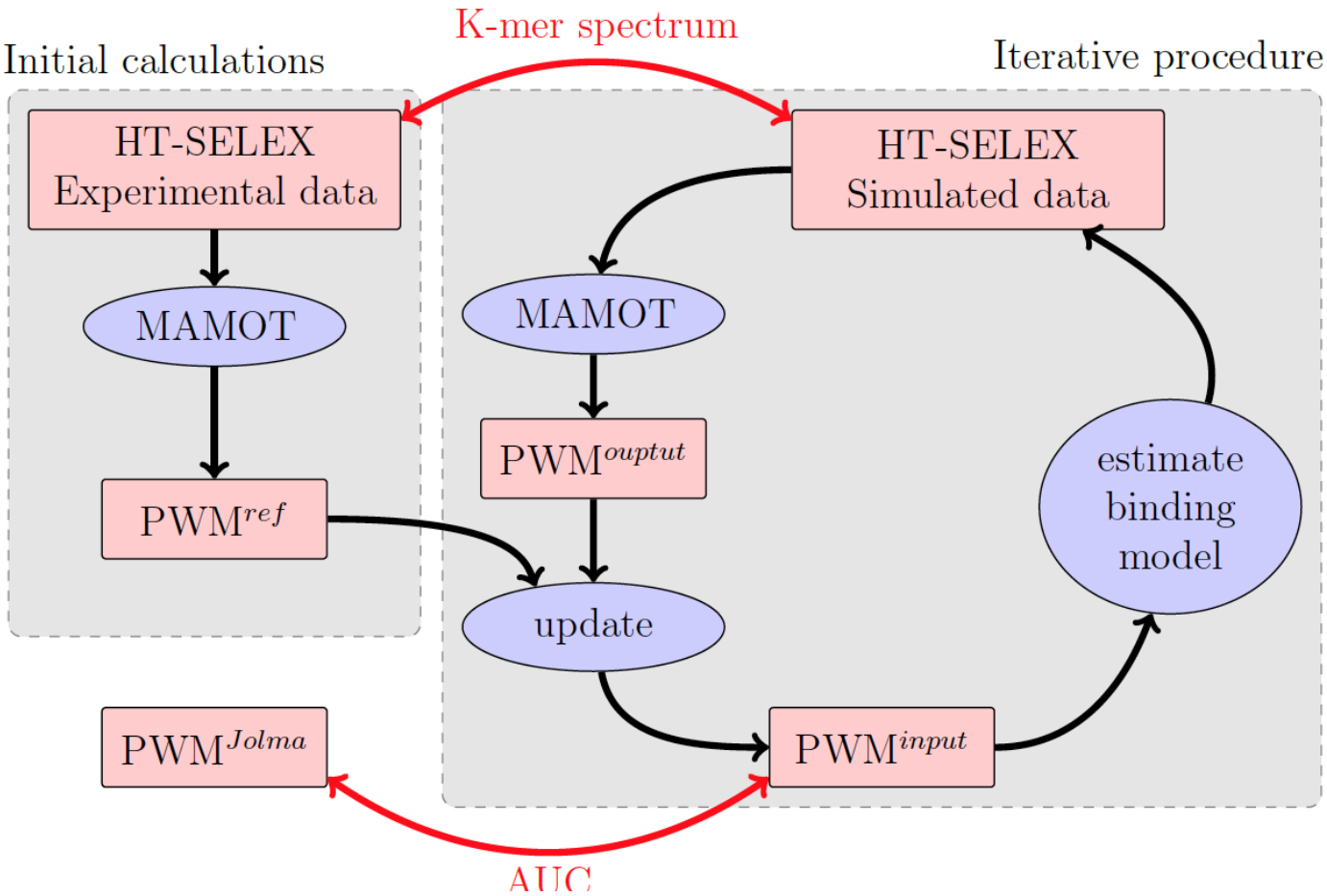
Weight matrix scores for successive CTF/NF1 HTP SELEX populations (Roulet et al. 2002)

high ← low affinity → high



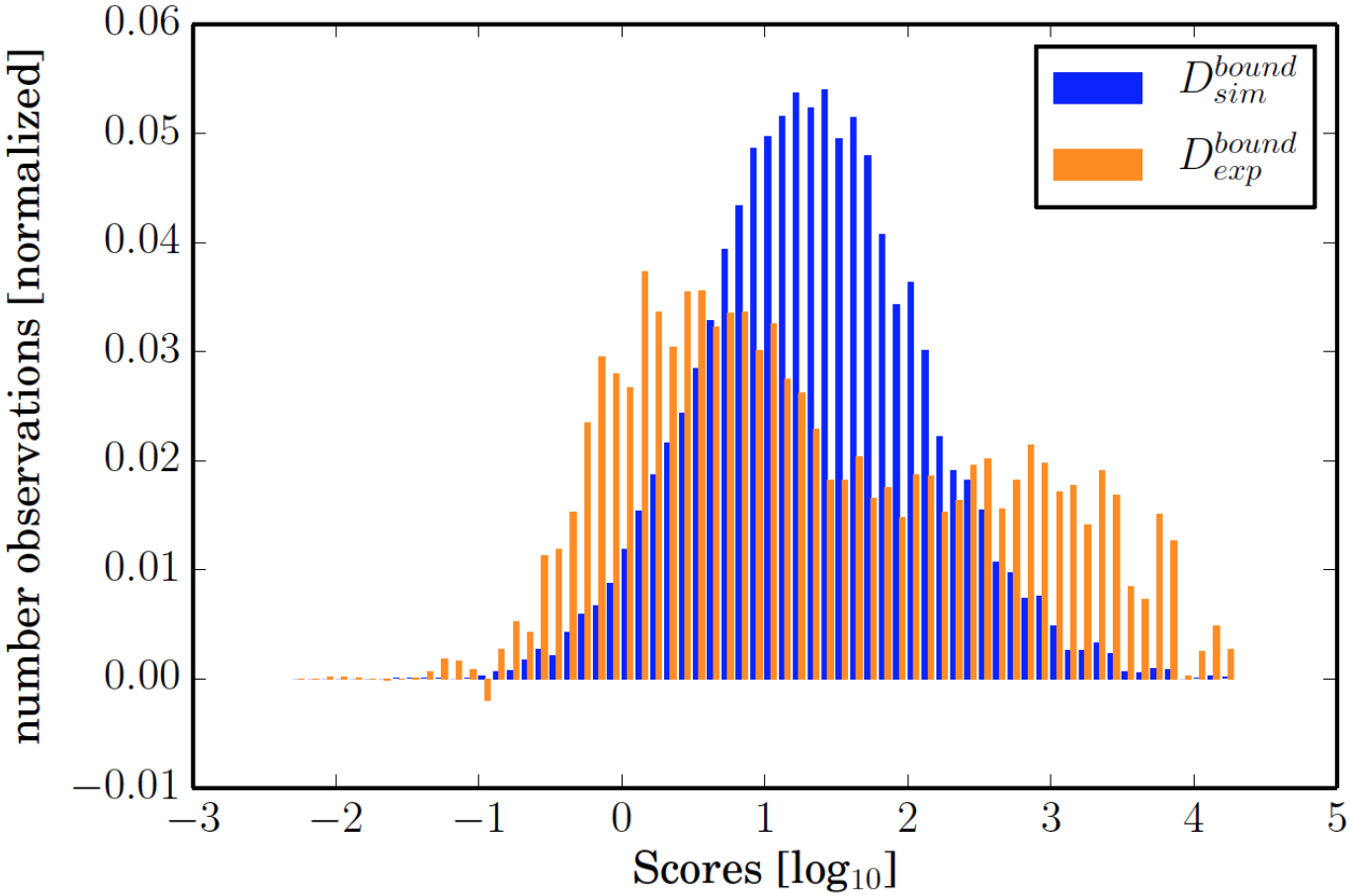


Reverse Simulations. Outline of a biophysically inspired-PWM inference procedure



Nicholas Molyneaux 2016, Master thesis.

First results: Theory doesn't always fit experiment data



Distributions of bound scores for the CEBPB protein. The experimental distribution shows two modes which cannot be reproduced by simulation based on the Berg & von Hippel model (Nicholas Molyneaux 2016, Master thesis.

Final remarks, Conclusions, Outlook

High-throughput SELEX data are great!

They potentially tell as a lot about the molecular mechanisms of sequence-specific DNA-protein interactions

The way we analysis them is far from optimal

Better TF specificity models could potentially be obtained with algorithms that take biophysical parameters of the selection process into account

Additional TF properties such as dimerization parameters and cooperative interaction parameter could also be learned with biophysical modeling

A lot more work to be done!

May Thanks to Two Very Important Scientific Collaborators:

Nicolas Mermod

For involving me in many exciting wet lab projects

Edgar Wingender

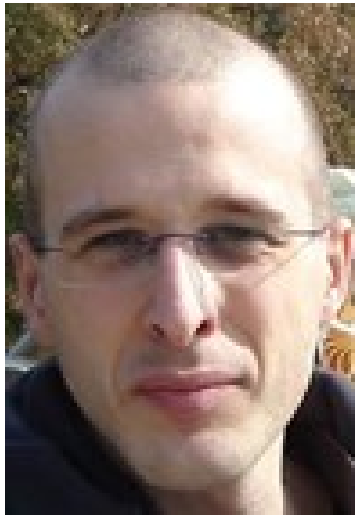
For long-lasting collaboration with the TRANSFAC team

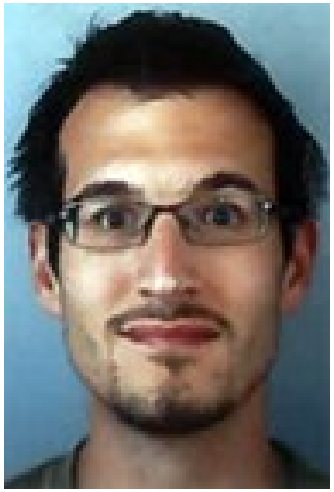
Many Thanks to my Current EPD Team Members

Rouayda Cavin Périer

Giovanna Ambrosini

René Dreos





Romain Groux